Log Concave Coupling for Sampling from Neural Net Posterior Distributions

Andrew R. Barron

YALE UNIVERSITY
Department of Statistics and Data Science
Joint work with Curtis McDonald

Singapore IMS-NUS Workshop on
Statistical Machine Learning for High-Dimensional Data
28 May 2024

Outline

- Neural Net Model and Approximation
  - Target f with variation $V_L(f)$ when represented with L layers
  - Approximation $f_{M,L}$ with L layers and M subnetworks
  - Approximation Accuracy $||f - f_{M,L}||^2 \leq \frac{V_L^2(f)}{M}$
- Neural Net Estimation and Risk
  - Estimate weights w, variation V, num subnets M, depth L
  - Constrained Least Squares: computational open problem
  - Bayes Predictive Mean Estimators: MCMC. Is it rapid?
  - Risk with sample size N and input dimension d

$$E||\hat{f} - f||^2 \leq \frac{V_L^2(f)}{M} + \frac{M \log(2d) + ML}{N}$$

$$E||\hat{f} - f||^2 \leq V_L(f) \sqrt{\frac{\log(2d) + L}{N}}$$

Outline: Continued

- Log Concave Coupling for Bayesian Computation
  - Focus attention on single hidden-layer network models
  - Prior density $p_0(w)$: Uniform on ℓ_1 constrained set
  - Posterior $p(w)$: Multimodal. No known direct rapid sampler
  - Coupling $p(\xi|w)$: cond indep Gaussian auxiliary variables $\xi_{i,m}$ with mean $x_i \cdot w_m$ for each observation i and neuron m
  - Conditional $p(w|\xi)$ always log-concave
  - Marginal $p(\xi)$ and its score $\nabla \log p(\xi)$ rapidly computable
  - $p(\xi)$ is log concave when the number of parameters Md is large compared to the sample size N
  - Langevin diffusion and other samplers are rapidly mixing
  - With a draw from $p(\xi)$ followed by a draw from $p(w|\xi)$ we obtain a draw from the desired posterior $p(w)$

Variation and Approximation with a Dictionary G

- Variation with respect to a dictionary
  - Dictionary G of functions $g(x, w)$, each bounded by 1
  - Consider linear combinations $\sum_j c_j\, g(x, w_j)$
  - Control the sum of abs values of the weights $\sum_j |c_j| \le V$
  - $\mathcal{F}_V =$ closure of signed convex hull of functions $V\, g(x, w)$
  - Variation $V_G(f) =$ the infimum of V such that $f \in \mathcal{F}_V$.
- Approximation accuracy
  - Function norm square $||f - g||^2$ in $L_2(P_X)$
  - M term approximation: $f_M(x) = \sum_{m=1}^{M} c_m\, g(x, w_m)$
  - Approximation error: $||f - f_M||^2 \le \frac{V(f)^2}{M}$
  - Trivial existence proof: Bernoulli, Hilbert, Maurey, Pisier, Barron 93
  - Greedy approximation proof: Jones, Barron 93
  - Outer weights c_m may equal $\pm\frac{V}{M}$
  - Approximation error better than $\frac{V^2}{M}$ is NP–hard (Vu 97)
  - Rate $\frac{1}{M}$ is dimension independent

Models

- Approximation error for $f_M(x) = \sum_{m=1}^{M} c_m \, g(x, w_m)$

$$||f - f_M||^2 \leq \frac{V_G^2(f)}{M}$$

- Algorithmic Terminology

  Sparse Term Selection, Variable Selection, Basis Selection, Forward Stepwise Regression, Relaxed Greedy Algorithm, Orthogonal Matching Pursuit, Frank Wolf Alg, Greedy Bayes
- Models

  Projection Pursuit (ridge functions), MARS (splines), MAPS (polynomials), Prony (sinusoids), Wavelets, Ridgelets, Random Forests (regression trees)
- Network Models

  Single Hidden-Layer Nets, Multi-Layer Networks, Deep Nets, Adaptive Learning Networks, Residual Networks
- Network Units (neurons)

  Sigmoids, Rectified Linear Units (ReLU), Polynomials, compositions thereof

Multi-Layer Neural Network Model

- Multi-Layer Net: Layers L, input x in $[-1,1]^d$, weights w
- Activation function: $\psi(z)$.
  - Rectified linear unit (ReLU): $\psi(z) = (z)_+$
  - Twice differentiable unit: sigmoid, smoothed ReLU, squared ReLU
- Paths of linked nodes: $\underline{j} = j_1, j_2, \ldots, j_L$.
- Path weight: $W_{\underline{j}} = w_{j_1,j_2} w_{j_2,j_3} \cdots w_{j_{L-1},j_L}$.
- Function representation:
  $f(x,c,w) = \sum_{j_L} c_{j_L} \psi\big(\sum_{j_{L-1}} w_{j_{L-1},j_L} \psi(\ldots\psi(\sum_{j_1} w_{j_1,j_2} x_{j_1})\ldots)\big)$
- Network Variation:
  - Internal: Sum abs. values of path weights set to 1.
  - External: $\sum_j |c_j| \leq V$
  - Variation: $V_L(f) =$ infimum of such V to represent f
  - Single Hidden-Layer Case: $V_1(f) \leq \int |\omega|_1^2 |\tilde{f}(\omega)| d\omega$ spectral norm
  - Class $\mathcal{F}_{L,V}$ of functions f with $V_L(f) \leq V$
- Interests: Approx, Metric Entropy, Stat. Risk, Computation

Complexity, Metric Entropy, Statistical Risk

- Gaussian complexity approach to bounding risk
  - Function class restricted to data:
    $\mathcal{F}^n = \{f(x_1), f(x_2), \ldots, f(x_n) : f \in \mathcal{F}\}$
  - Gaussian Complexity:
    $C(A) = (1/\sqrt{n}) E_Z[\sup_{a \in A} a \cdot Z]$ for $Z \sim N(0, I)$ $A \subset R^n$
  - Complexity of Neural Nets:
    $C(\mathcal{F}^n_{L,V}) \leq V\sqrt{2 \log 2d + 2L \log 2}$
    for ψ Lipshitz 1 via Fernique Gaussian comparison ineq,
    Klusowski, B. 2020 (cf Neshabur et al 15, Golowich et al 18)
  - Gaussian complexity provides control of
    - Metric Entropy:
      $\log |Cover(\mathcal{F}_{L,V}, \delta)| \leq \frac{16 C^2(\mathcal{F}_{L,V})}{\delta^2}$
    - Stat Risk of Constrained Least Squares:
      $E||\hat{f} - f||^2 \leq \frac{8 C(\mathcal{F}_{L,V})}{\sqrt{n}}$

Minimum Description Length and Bayes predictive risk

- Minimum Description Length; optimize penalized likelihood
  - Least squares with suitable penalization for choice of M, V
  - $||\hat{f} - f||^2$ risk via Renyi-Battacharya risk inequality: B, Luo 08
  - Index of Resolvability: ApproxError + Complexity/N
- Predictive Bayes and its cumulative risk control
  - Predictive density $\hat{p}_n(y|x) = \int p(y|x,w)p(w|x^n,y^n)dw$
  - Predictive mean $\hat{f}_n(x) = \int f(x,w)p(w|x^n,y^n)dw$
  - Predictive evaluations for $Y_{n+1} = y$ when $X_{n+1} = x$
  - Inf Thy chain rule for cumulative Kullback risk: B. 86,98

    $$\frac{1}{N}\sum_{n=0}^{N-1} ED(P^*_{Y|X}||\hat{P}^n_{Y|X}) = \frac{1}{N}D(P^*_{Y^N,X^N}||P_{Y^N,X^N})$$
  - Controls data compression redundancy as well as the risk
  - Index of Resolvability:

    ApproxError $+\ \frac{1}{N}$ log[1 / PriorProb(ApproxSet)]
  - Used in Yang, B (98) minimax risk characterization

    $$E||f - \hat{f}_N||^2 \leq \min_\delta \left\{\delta^2 + \frac{1}{N}\log|\mathrm{Cover}(\mathcal{F}_{L,V},\delta)|\right\} \leq \frac{8C(\mathcal{F}_{L,V})}{\sqrt{N}}$$

Arbitrary Sequence Predictive Bayes Regret

- On-line learning
- Arbitrary-sequence regret for predictive Bayes

$$\frac{1}{N}\sum_{n=1}^{N}(Y_n - \hat{f}_{n-1}(X_n))^2 - \frac{1}{N}\sum_{n=1}^{N}(Y_n - f(X_n))^2$$

- Bound hold of the same form, uniformly over X^N, Y^N,

$$Regret_N \le Approx\ Error + \frac{1}{N}\log\frac{1}{PriorProb(Approx\ Set)}$$

- Specialization of bound to the case of functions f in $F_{1,V}$

$$Regret_N \le V\frac{\sqrt{logd}}{\sqrt{N}}$$

- Taking expectation controls

$$\frac{1}{N}\sum_{n=1}^{N} E||f - \hat{f}_{n-1}||^2$$

- Estimator $\hat{\hat{f}}_N(x) = \frac{1}{N}\sum_{n=1}^{N}\hat{f}_{n-1}$ also has this bound

$$E[||\hat{\hat{f}}_N - f||^2] \le V\frac{\sqrt{logd}}{\sqrt{N}}$$

Bayesian Computation for Neural Net

- Sample sizes: $n \le N$
- $\text{Data}^n = ((X_i, Y_i) \text{ for } i = 1, 2, \ldots, n)$, with X_i in $[-1, 1]^d$
- Natural yet optional statistical assumption:
  (X_i, Y_i) independent $P_{X,Y}$, with target $f(x) = E[Y \mid X = x]$
  - Not needed for Bayesian computation statements
  - Not needed for online learning bounds
- Single hidden-layer network model: $f(x, \underline{w})$
  $f_M(x, \underline{w}_1, \ldots \underline{w}_M) = \frac{V}{M} \sum_{m=1}^{M} \psi(\underline{w}_m \cdot x_i)$
- each $\underline{w}_m$ in symmetric simplex $S_1^d = \{w : \sum_{j=1}^{d} |w_j| \le 1\}$
- Prior: $p_0(\underline{w})$ makes $\underline{w}_m$ independent uniform on S_1^d
- Likelihood: $\exp\{-\beta g(w)\}$
  where $g(w) = \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \frac{V}{M} \sum_{m=1}^{M} \psi(x_i \cdot w_m)\right)^2$
- Posterior: $p(w) = p_0(w) \exp\{-\beta g(w) - \Gamma(\beta)\}$
- Bayesian Computation: Estimate $\hat{f}(x) = \int f(x, w) p(w) dw$
  by drawing independent samples from $p(w)$ and averaging $f(x, w)$

Hessian of the Minus Log Likelihood

- Log 1/Likelihood $= \beta\, g(w)$
  - Gradient $\text{score}(w) = \beta\, \nabla g(w)$
  - Hessian $= \beta H(w) = \beta\, \nabla\nabla' g(w)$
- Squared error loss: $g(w)$
  $\frac{1}{2}\sum_{i=1}^{n}(res_i(w))^2$ where $res_i(w) = Y_i - \frac{V}{M}\sum_{m=1}^{M}\psi(x_i \cdot w_m)$
- Gradient: $\nabla_{w_m} g(w)$ for block m
  $$-\frac{V}{M}\sum_{i=1}^{n} res_i(w)\psi'(x_i \cdot w_m)x_i$$
- Hessian: $H_{w_k,w_m}(w) = \nabla_{w_k}\nabla'_{w_m} g(w)$ for block k, m
  $$\frac{V^2}{M^2}\sum_{i=1}^{n}\psi'(x_i \cdot w_k)\psi'(x_i \cdot w_m)x_i x_i'$$
  $$-\frac{V}{M}\sum_{i=1}^{n} res_i(w)\psi''(x_i \cdot w_m)x_i x_i' 1_{k=m}$$
- Quadratic form: $a'H(w)a$, where a has blocks a_m $1 \le m \le M$
  $$\frac{V^2}{M^2}\sum_{i=1}^{n}\left(\sum_{m=1}^{M}\psi'(x_i \cdot w_m)\, a_m \cdot x_i\right)^2$$
  $$-\frac{V}{M}\sum_{i=1}^{n} res_i(w)\sum_{m=1}^{M}\psi''(x_i \cdot w_m)(a_m \cdot x_i)^2$$
- $p(w)$ is not log-concave; that is, $g(w)$ is not convex
  The first term is positive definite, the second term is not
- No clear reason for gradient methods to be effective

Log Concave Coupling

- Auxiliary Random Variables $\xi_{i,m}$ chosen conditionally indep
- Normal with mean $x_i \cdot w_m$, variance $1/\rho$, with $\rho = \beta c V/M$
- Conditional density:

  $$p(\xi|w) = (\rho/2\pi)^{Mn/2} exp\{-\frac{\rho}{2}\sum_{i=1}^{n}\sum_{m=1}^{M}(\xi_{i,m} - x_i \cdot w_m)^2\}$$
- Multiplier $c = c_{Y,V} = \max_i |Y_i| + V$ exceeds $|res_i(w)|$ for all w
- Activation second derivative: $|\psi''(z)| \leq 1$ for $|z| \leq 1$
- Joint density: $p(w,\xi) = p(w)p(\xi|w)$
- Reverse conditional density:

  $$p(w|\xi) = p_0(w)\exp\{-\beta g_\xi(w) - \Gamma_\xi(\beta)\}$$
- Conditional log 1/Likelihood $= \beta g_\xi(w)$ with

  $$g_\xi(w) = g(w) + \frac{1}{2}\frac{V}{M}c\sum_{i=1}^{n}\sum_{m=1}^{M}\left(x_i \cdot w_m - \xi_{i,m}\right)^2$$
- Modifies Hessian $a' H_\xi(w) a$ with new positive def second term

  $$\frac{V}{M}\sum_i\sum_m [c + res_i(w)\psi''(x_i \cdot w_m)](a_m \cdot x_i)^2$$
- $p(w|\xi)$ is log concave in w for each ξ
- Efficiently sample. MCMC theory, Lovasz, Kannan, Vempala,...

Marginal Density and Score of the Auxiliary Variables

- Auxiliary variable density function:
  $p(\xi) = \int p(w, \xi) dw$
  Integral of a log concave function of w
- Rule for Marginal Score:
  $\nabla \log 1/p(\xi) = E[\nabla \log 1/p(\xi|w) \,|\, \xi\,]$
- Normal Score: linear
  $\partial_{\xi_{i,m}} \log 1/p(\xi|w) = \rho \left(\xi_{i,m} - x_i \cdot w_m\right)$
- Marginal Score:
  $\partial_{\xi_{i,m}} \log 1/p(\xi) = \rho \left(\xi_{i,m} - E[x_i \cdot w_m \,|\, \xi\,]\right)$
- Efficiently compute ξ score by Monte Carlo sampling of $w|\xi$
- Permits Langevin stochastic diffusion: with gradient drift
  $d\,\xi(t) = \frac{1}{2} \nabla \log p(\xi(t))\, dt + d\,B(t)$
  converging to a draw from the invariant density $p(\xi)$
- Lyapunov function identification $e^{\alpha||\xi||^2}$ as in Hairer (21) reveals exponential convergence $||p_t - p||_1 \le e^{-t/\tau}$
- What is the size of $\tau > 0$?

Hessian of $\log 1/p(\xi)$. Is $p(\xi)$ log concave?

- Hessian of $\log 1/p(\xi)$, an nM by nM matrix

$$\tilde{H}(\xi) = \nabla\nabla' \log 1/p(\xi) = \rho \left\{ I - \rho\, Cov \left[\begin{matrix} Xw_1 \\ \cdots \\ Xw_M \end{matrix} \,\middle|\, \xi \right] \right\}$$

- Hessian quadratic form for unit vectors a in R^{nM} with blocks a_m

$$a'\tilde{H}(\xi)a = \rho\,\{1 - \rho\, Var[\tilde{a}\cdot w|\xi]\}$$

where $\tilde{a} = \begin{bmatrix} X'a_1 \\ \cdot \\ X'a_M \end{bmatrix}$ has $||\tilde{a}||^2 \le n\, d$

- Requires variance of $\tilde{a}\cdot w$ using the log-concave $p_\beta(w|\xi)$
- More concentrated, smaller variance, than with the prior?
- Counterpart using the prior

$$\rho\,\{1 - \rho\, Var_0[\tilde{a}\cdot w]\}$$

- Use $Cov_0(w_m) = \frac{2}{(d+2)(d+1)} I$ and $\rho = \beta c V/M$ to see its at least

$$\rho \left\{ 1 - \frac{2\beta c V n}{M(d+2)} \right\}$$

- Constant β chosen such that $\beta c V \le 1/4$
- Strictly positive when number param Md exceeds sample size n
- Hessian $\ge (\rho/2) I$. Strictly log concave

Rapid Convergence of Stochastic Diffusion

- Recall the Langevin diffusion

$$d\,\xi(t) = \tfrac{1}{2}\nabla \log p(\xi(t)dt \,+\, d\,B(t)$$

- There are time-discretizations (e.g. Metropolis adjusted)
- Natural initialization choice $\xi(0)$ distributed $N(0,(1/\rho)I)$
- Bakry-Emery theory (initiated in 85)
- Strong log concavity yields rapid Markov proc. convergence
- In particular, in the stochastic diffusion setting

$$\nabla\nabla' \log 1/p(\xi) \geq (\rho/2)I$$

  yields exponential conv. of relative entropy (Kullback distance)

$$D(p_t||p) \leq e^{-t\rho/2} D_0$$

- In particular, the time required for small relative entropy is controlled by $\tau = 2/\rho$, here equal to $2M/(\beta c V)$

Is $p(\xi)$ log concave?

- Recap: quadratic form in Hessian of $\log 1/p(\xi)$
  $$a'\tilde{H}(\xi)a = \rho\{1 - \rho\, Var[\tilde{a} \cdot w|\xi]\}$$
- Suppose $p(w|\xi)$ has smaller variance than with the prior. Then, when Md exceeds n, this Hessian is strictly positive def, that is, $p(\xi)$ is strictly log concave
- Other available controls on the variance
  $$Var[\tilde{a} \cdot w|\xi] \le \int (\tilde{a} \cdot w)^2 \exp\{-\beta g_\xi(w) - \Gamma_\xi(\beta)\} p_0(w)dw$$
- Hölder's inequality
  $$< [E_0[(\tilde{a} \cdot w)^{2k}]]^{1/k} \exp\{\tfrac{k-1}{k}\Gamma_\xi(\tfrac{k}{k-1}\beta) - \Gamma_\xi(\beta)\}$$
- Deduce a choice of k such that $\rho Var[\tilde{a} \cdot w|\xi]$ is less than 1, for all Md at least a suitable power of n
- Carried out this agenda in a related greedy Bayes model (next)

Greedy Bayes

- Initialize $\hat{f}_{n,0}(x) = 0$
- Given previous neuron fits, iterate k, for each n

$$f_{n,k}(x, w) = (1-\alpha) f_{n,k-1}(x) + \lambda\psi(w \cdot x)$$

- $\alpha = 1/\sqrt{N}$ and $\lambda = V\alpha$ are suitable.
- Form the iterative squared error $g(w)$

$$g_{n,k}(w) = \frac{1}{2}\sum_{i=1}^{n-1}(y_i - f_{i,k}(x_i, w))^2$$

  Again Hessian has a not necessarily positive definite part

$$-\lambda \sum_{i=1}^{n-1} r_{i,k-1}\psi''(w \cdot x_i) x_i x_i'$$

- Associated greedy posterior $p_{n,k}(w)$ proportional to

$$p_0(w) \exp\{-\beta g_{n,k}(w)\}$$

- Update $f_{n,k}$ replacing $\psi(w \cdot x)$ with its posterior mean
- Estimate by sampling from the greedy posterior

Log Concave Coupling for Greedy Bayes

- For the moment, fix n, k
- Again $p(w) = p_0(w) \exp\{-\beta g(w)\}$
- Coupling random variables $\xi_i \sim N(x_i \cdot w, 1/\rho)$ with $\rho = c\lambda\beta$
- Joint density $p(w, \xi)$ with logarithm $-\beta g(w, \xi)$ built from
  $$\beta g(w) + \frac{1}{2} c\lambda \sum_{i=1}^{n-1} (\xi_i - w \cdot x_i)^2$$
  which is convex in w for each ξ, so $p(w|\xi)$ is log concave
- The associated marginal is $p(\xi)$
- Hessian quadratic form $a' \nabla\nabla \log 1/p(\xi) a$
  $$\rho\{1 - \rho Var[\tilde{a} \cdot w|\xi]\}$$
  for a with $||a|| = 1$ and $\tilde{a} = X' a$
- Deduce $p(\xi)$ is log concave for sufficiently large d
- From which get w by a draw from $p(w|\xi)$

Variance control using Hölder's inequality

- As before $Var[\tilde{a} \cdot w|\xi]$ is not more than
  $$\int (\tilde{a} \cdot w)^2 \exp\{-\beta g_\xi(w) - \Gamma_\xi(\beta)\}\, p_0(w)\, dw$$
  where $g_\xi(w)$ is $g(w, \xi)$ minus its mean value at $\beta = 0$
- $\Gamma_\xi(w)$ is the cumulant generating function of $-g_\xi(w)$
- By Hölders inequality that variance is not more than
  $$[E_0[(\tilde{a} \cdot w)^{2k}]]^{1/k} \exp\{\tfrac{k-1}{k}\Gamma_\xi(\tfrac{k}{k-1}\beta) - \Gamma_\xi(\beta)\}$$
- For the first factor, for unit vectors ν
  $$E_0[(\nu \cdot w)^{2k}] \leq \frac{(2k)!}{(d+2k)\cdots(d+1)}$$
- Implication
  $$[E_0[(\tilde{a} \cdot w)^{2k}]]^{1/k} \leq ||\tilde{a}||^2 \left(\frac{2k}{ed}\right)^2$$

On the second factor from Hölders inequality

- The exponent of the second factor is

$$\frac{k-1}{k}\Gamma_\xi\left(\frac{k}{k-1}\beta\right) - \Gamma_\xi(\beta)$$

- It takes the form $\beta^2 Var_{\tilde{\beta}}[g_\xi(w)|u]/(k-1)$
- Subtracting the value at 0 we have for some $\tilde{w}$

$$g_\xi(w) - g_\xi(0) = w'\nabla g_\xi(\tilde{w})$$

- So the max square of $w'\nabla g_\xi(\tilde{w})$ bounds variance of $g_\xi(w)$
- Indeed a value near $(2c\lambda n)^2$ bounds that variance
- Optional page verifies this for a suitable set of u
- Hence the exponent of second factor not more than value near

$$4\beta^2\lambda^2c^2n^2/k$$

Optional page verifying bound on $w'\nabla g_\xi(\tilde{w})$

- Concerning $\nabla g_\xi(\tilde{w})$ it is
$$-\lambda\left\{\sum_{i=1}^{n-1}\left[res_{i,k-1}\psi'(\tilde{w}\cdot x_i) - c\tilde{w}\cdot x_i\right] x_i + \sum_{i=1}^{n-1}\xi_i x_i\right\}$$
- Hit with w, the result has magnitude not more than
$$2c\lambda n + \lambda \max_j \left|\sum_{i=1}^{n-1}\xi_i x_{i,j}\right|$$
- With high probability, the max is not more than $\kappa\sqrt{n/\rho}$ where $\kappa = \sqrt{2\log 2d}$
- So the max is of smaller order than the first term
- Restrict the ξ to have such bound
- Then exponent of second factor not more than value near
$$4\beta^2\lambda^2 c^2 n^2/k$$

Combining the two factors

- Use $\tilde{a} = \sum_i a_i x_i$ with $||\tilde{a}||^2 \leq nd$ and $\rho = c\lambda\beta$
- Combine the two factors
- Obtain $\rho Var[\tilde{a} \cdot w|\xi]$ not more than
  $$c\lambda\beta \, nd(2k/(ed)^2 \, exp\{4\beta^2\lambda^2 c^2 n^2/k\}$$
- The optimal $k = 2\beta^2\lambda^2 c^2 n^2$ yielding not more than
  $$4(c\lambda\beta n)^5/d$$
- Recall $\lambda = V\alpha = V/\sqrt{n}$
- Choose $\beta = 1/(2cV)$, choose $d \geq n^{5/2}$.
- $\rho Var[\tilde{a} \cdot w|\xi]$ is strictly less than 1 indeed (less than $1/2$)
- Hence $p(\xi)$ is strictly log concave, for d exceeding $n^{5/2}$

Summary

- Multimodal neural net posteriors can be efficiently sampled
- Log concave coupling provides the key trick
- Requires number of parameters Md large compared to the sample size N
- Statistically accurate provided ℓ_1 controls are maintained on the parameters
- Provides the first demonstration that the class $\mathcal{F}_{1,V}$ associated with single hidden layer networks (including the class of functions with bounded L_1 spectral norm) is both computationally and statistically learnable
- A polynomial number of computations in the size of the problem is sufficient
- The approximation rate $1/M$ and the statistical learning rate $1/\sqrt{N}$ are independent of the dimension for this class of functions